PAC learning

ncordon

Índice

Estos apuntes son una adptación en su mayoría del contenido del libro [**?**]

1. Introducción

Damos unas notaciones/definiciones básicas que utilizaremos de aquí en adelante.

- **Dominio**: $\mathcal{X}$, sobre el que tenemos definida una σ álgebra de conjuntos B. Llamamos una instancia a $x \in \mathcal{X}$
- **Conjunto de etiquetas**: $\mathcal{Y} \subseteq \mathbb{R}$ finito , que asumiremos como $\{0, 1\}$ en lo que sigue hasta que se indique lo contrario. Esto nos restringe al paradigma de clasificación binario.
- **Verdadero etiquetado**: Asumimos la existencia de una función $f : \mathcal{X} \to \mathcal{Y}$ que devuelve el verdadero etiquetado de todas las instancias.
- **Generación de instancias**: Asumimos la existencia de una distribución de probabilidad $\mathcal{D}$ sobre $\mathcal{X}$, para la σ álgebra de conjuntos mencionada anteriormente, que nos da información sobre la probabilidad de extraer cada posible instancia desde $\mathcal{X}$.
- **Conjunto de entrenamiento**: Tenemos una muestra aleatoria simple $S = (\mathcal{X}_1, \ldots, \mathcal{X}_m)$, idéntica e independientemente distribuida, donde $S \sim \mathcal{D}^m$, esto es cada X_i sigue la misma distribución que $\mathcal{X}$, $X_i \sim \mathcal{D}$, y las distribuciones marginales son independientes entre sí. Notaremos S_x a una realización muestral $(x_1, \ldots x_m)$. Cada elemento x_i de una realización muestral $S_x = (x_1, \ldots x_n)$ se etiqueta por f, y llamando $f(x_i) = y_i$ definimos como conjunto de entrenamiento a la tupla $((x_1, y_1), \ldots, (x_m, y_m))$. La relación entre la realización muestral y el conjunto de entrenamiento asociado es biunívoca, por lo que por abuso de notación llamaremos indiferentemente conjunto de entrenamiento a ambas tuplas.
- **Resultado del aprendizaje**: disponemos de un algoritmo de aprendizaje $A : (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{Y}^{\mathcal{X}}$ que recibe un conjunto de entrenamiento y devuelve una función $h : \mathcal{X} \to \mathcal{Y}$ que llamaremos hipótesis/clasificador. El algoritmo "desconoce.[el] valor de la verdadera función de etiquetado f en los puntos no pertenecientes al conjunto de entrenamiento.
- **Error del clasificador**: Definimos el error de un clasificador $h : \mathcal{X} \to \mathcal{Y}$ como:

$$L_{\mathcal{D},f}(h) := P(\{x \in \mathcal{X} : h(x) \neq f(x)\}) = P[f \neq h]$$

Por simplificar la escritura, omitiremos a partir de ahora el hecho de que sobre $\mathcal{X}$ tenemos definida una σ álgebra de conjuntos, B, y que todas las distribuciones asignan probabilidad a los conjuntos de alguna σ álgebra que contenga a B. Además, consideraremos que la función de verdadero etiquetado y los clasificadores son funciones medibles para que la definición de los errores sea correcta.

1.1. Minimización del riesgo empírico (ERM)

Definición 1. ***Riesgo empírico (ER)***

Fijado un clasificador $h : \mathcal{X} \to \mathcal{Y}$, definimos el riesgo empírico o error empírico, como una variable aleatoria:

$$\begin{array}{rccl} L_S(h): & \mathcal{X}_1 \times \ldots \times \mathcal{X}_m & \to & \mathbb{R} \\ & (x_1, \ldots x_m) & \mapsto & L_{S_x}(h) = \dfrac{\#\{i \in 1 \ldots m : h(x_i) \neq f(x_i)\}}{m} \end{array}$$

Para un conjunto de entrenamiento el riesgo empírico proporciona el error del clasificador sobre el conjunto de entrenamiento.

Un algoritmo que obtiene una hipótesis que minimiza el error empírico sobre un conjunto de entrenamiento recibe el nombre de ERM y notamos $ERM(S_x)$ al clasificador obtenido con dicho algoritmo.

Este error no es siempre óptimo. Pensemos en el siguiente ejemplo:

Sea $\mathcal{X} = \mathbb{R}$, $\mathcal{D}$ la distribución uniforme sobre $[0, 2] \subset \mathbb{R}$, y la siguiente función:

$$f(x) = \begin{cases} 1 & x \in [0, 1] \\ 0 & x \in \mathbb{R} \setminus [0, 1] \end{cases}$$

Sea $((x_1, y_1), \ldots (x_m, y_m))$ un conjunto de entrenamiento de tamaño m y el clasificador:

$$h(x) = \begin{cases} y_i & \exists i \in \{1 \dots m\} : x = x_i \\ 0 & \nexists i \in \{1 \dots m\} : x = x_i \end{cases}$$

Nótese que el conjunto de entrenamiento no puede tener elementos no repetidos puesto que se etiquetan mediante f, que es una función y no puede arrojar dos imágenes distintas para un mismo $x \in \mathcal{X}$ de entrada.

Este clasificador es perfecto respecto a la minimización de riesgo empírico, pero $L_{\mathcal{D},f}(h) = 1/2$. Es decir, tiene el mismo nivel de acierto que el clasificador idénticamente 1. A este fenómeno, minimizar el riesgo empírico siendo un clasificador con un error muy alto, lo denominamos *overfitting*.

El hecho de tomar el error sobre el conjunto de entrenamiento como aproximación al verdadero error del clasificador se respalda por la siguiente proposición:

Proposición 1. ***Relación entre riesgo empírico y error del clasificador***

Sea $\mathcal{H}$ clase de clasificadores binarios sobre un dominio $\mathcal{X}$. Sea $\mathcal{D}$ una distribución desconocida sobre $\mathcal{X}$. Sea f la función de verdadero etiquetado. Para $h \in \mathcal{H}$ fijo se verifica:

$$\mathbb{E}[L_S(h)] = L_{\mathcal{D},f}(h)$$

Llamamos $p = P[f \neq h] = L_{\mathcal{D},f}(h)$

$$\begin{aligned} \mathbb{E}[L_S(h)] &= \sum_{k=0}^{m} \frac{k}{m} \binom{m}{k} p^k (1-p)^{m-k} = \sum_{k=1}^{m} \frac{k}{m} \binom{m}{k} p^k (1-p)^{m-k} = \\ &= \sum_{k=1}^{m} \binom{m-1}{k-1} p^k (1-p)^{m-k} = \sum_{k=0}^{m-1} \binom{m-1}{k} p^{k+1} (1-p)^{m-1-k} = \\ &= p \cdot \sum_{k=0}^{m-1} \binom{m-1}{k} p^k (1-p)^{m-1-k} = p(1+(1-p))^{m-1} = p \end{aligned}$$

1.2. ERM con *sesgo inductivo*

Con objeto de corregir el ERM, para evitar *overfitting*, usamos el conocimiento previo sobre el problema (la información que dispongamos sobre el dominio, la distribución, etc) restringiendo el espacio de búsqueda, esto es, la clase de hipótesis $\mathcal{H}$ desde la que el algoritmo puede escoger un $h : \mathcal{X} \to \mathcal{Y}$. Llamamos a esto **sesgo inductivo** puesto que se asumirá una determinada clase de funciones $\mathcal{H}$ en función de las características del problema.

Notaremos a un clasificador obtenido con este paradigma $h_{S_x} := ERM_{\mathcal{H}}(S_x)$, y lo definimos de manera que:

$$h_{S_x} \in \underset{h \in \mathcal{H}}{argmin}\{L_{S_x}(h)\}$$

La existencia de $\underset{h \in \mathcal{H}}{min}\{L_{S_x}(h)\}$ está garantizada, ya que $m \cdot L_{S_x}(h) \in \mathbb{N}$ para todo $h \in \mathcal{H}$.

Enunciamos la propiedad de factibilidad, que usaremos más adelante.

Definición 2. ***Propiedad de factibilidad***

Existe $\bar{h} \in \mathcal{H}$ *verificando* $L_{\mathcal{D},f}(\bar{h}) = 0$.

La hipótesis de factibilidad implica que $P[L_S(\bar{h}) = 0] = 1$, ya que:

$$\begin{aligned} P(\{(x_1, \dots x_m) : \bar{h}(x_i) = f(x_i), i = 1, \dots m\}) = \\ = \prod_{i=1}^{m} P[h = f] = \prod_{i=1}^{m} (1 - P[h \neq f]) = 1 \end{aligned}$$

Por tanto $P[L_S(h_S) = 0] = 1$.

Para finalizar estos preliminares remarcamos que el valor de $L_{\mathcal{D},f}(h_{S_x})$ dependerá del conjunto de entrenamiento, extraído y etiquetado a partir del vector aleatorio S, y la elección del mismo está sometida al azar. Asimismo, necesitamos una medida de la bondad de la predicción.

2. Aprendizaje PAC.

Definición 3. ***PAC (Probablemente Aproximadamente Correcto) cognoscible***

Una clase de funciones $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ es PAC cognoscible sii existe una función $m_{\mathcal{H}} :]0,1[^2 \to \mathbb{N}$, llamada complejidad muestral, y un algoritmo A verificando que si $0 < \epsilon, \delta < 1$, entonces para toda distribución $\mathcal{D}$ sobre $\mathcal{X}$ y para toda función de verdadero etiquetado $f : \mathcal{X} \to \{0,1\}$ cumpliendo la propiedad de factibilidad, ejecutando el algoritmo para un conjunto de entrenamiento generado por $S \sim \mathcal{D}^m$ etiquetado mediante f, con $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ se tiene que:

$$P[L_{\mathcal{D},f}(A(S)) \leq \epsilon] \geq 1 - \delta$$

Llamamos a $(1-\delta)$ *confianza de la predicción* y a $(1-\epsilon)$ la *exactitud*. Estos dos parámetros explican el nombre aproximadamente ($\leftrightarrow$ confianza) correcto ($\leftrightarrow$ exactitud).

Podemos considerar $m_{\mathcal{H}}$ única en el sentido de que para cada (δ, ϵ) nos devuelva el menor natural verificando las hipótesis del enunciado.

Nótese que las condiciones exigidas, cumplir la propiedad de factibilidad y que la hipótesis devuelta deba estar en $\mathcal{H}$, son muy fuertes. Relajaremos esta definición más adelante con el concepto de PAC agnóstico.

2.1. Aprendizaje con clases finitas

Teorema 1. ***Las clases finitas de funciones son PAC cognoscibles***

Sea $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ finito. Sean $0 < \epsilon, \delta < 1$, y un natural $m \in \mathbb{N}$ verificando:

$$m \geq \frac{1}{\epsilon} log\left(\frac{|\mathcal{H}|}{\delta}\right)$$

Entonces para toda función de verdadero etiquetado $f : \mathcal{X} \to \{0,1\}$, y para toda distribución $\mathcal{X} \sim \mathcal{D}$ para la que se verifique la propiedad de factibili-

dad, las hipótesis que obtenemos a través del algoritmo ERM son con una confianza superior a $1-\delta$, $1-\epsilon$ *exactas.*

Como consecuencia, deducimos que la complejidad muestral es menor o igual a $\left\lceil \frac{1}{\epsilon} log\left(\frac{|\mathcal{H}|}{\delta}\right)\right\rceil$

Demostración. Fijada una distribución $\mathcal{D}$, $m \in \mathbb{N}$ y una función de etiquetado f, notamos:

$$\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}$$

Se tiene:

$$P[L_{\mathcal{D},f}(h_S) > \epsilon] \leq P[\exists h \in \mathcal{H}_B : L_S(h) = 0] \leq \sum_{h \in \mathcal{H}_B} P[L_S(h) = 0]$$

La primera desigualdad viene dada porque dada h_{S_x} se verifica, por la propiedad de factibilidad, que $L_S(h_{S_x}) = 0$. La segunda por subaditividad.

Además, fijada $h \in \mathcal{H}_B$, como $L_{\mathcal{D},f}(h) > \epsilon$:

$$\begin{aligned} P[L_S(h) = 0] &= P(\{(x_1, \dots x_m) : h(x_i) = f(x_i), i = 1, \dots m\}) = \\ &= \prod_{i=1}^{m} P[h = f] = \prod_{i=1}^{m} (1 - L_{\mathcal{D},f}(h)) \leq (1-\epsilon)^m \leq e^{-\epsilon m} \end{aligned}$$

Las dos desigualdades probadas, junto a la hipótesis del enunciado, y usando $\mathcal{H}_B \subseteq \mathcal{H}$ dan lugar a:

$$P_{S \sim \mathcal{D}^m}[L_{\mathcal{D},f}(h_S) > \epsilon] \leq |\mathcal{H}| e^{-\epsilon m} \leq \delta$$

□

2.2. Aprendizaje con clases no finitas

¿Hay ejemplos de clases infinitas PAC cognoscibles? Veamos un ejemplo.

Definición 4. ***Clasificadores de rectángulo***

Un clasificador de rectángulo es un clasificador que asigna 1 a los puntos que se quedan dentro de un cierto rectángulo en el plano real.

$$h_{a,b,c,d} = \mathbb{1}_{[a,b]\times[c,d]}$$

La clase de clasificadores de rectángulo en el plano se define por:

$$\mathcal{H}^2_{rec} = \{h_{a,b,c,d} : a \leq b, c \leq d\}$$

Proposición 2. ***Los rectángulos son PAC cognoscibles***

Asumiendo propiedad de factibilidad, los rectángulos son PAC cognoscibles

Sea A el algoritmo que devuelve el rectángulo más pequeño que engloba a todos los ejemplos positivos del conjunto de entrenamiento S_x.

Partiendo de la propiedad de factibilidad, debe existir un clasificador de rectángulo $\bar{h} = h_{a,b,c,d}$ que haga el ERM nulo y que cumpla $L_{\mathcal{D},f}(\bar{h}) = 0$. Por tanto debe verificarse que h_{S_x} debe acertar en todas las instancias positivas (cuya etiqueta sea 1) del conjunto de entrenamiento, con probabilidad 1, ya que si valiese 0 en algún ejemplo positivo del conjunto de entrenamiento, el ERM sería mayor que 0.

El algoritmo que devuelve el mínimo rectángulo que engloba a todos los ejemplos positivos es por tanto un ERM.

Veamos que con este algoritmo minimizador del ERM la clase de rectángulos es PAC cognoscible.

Sea $R^* = [a, b] \times [c, d]$ el rectángulo que materializa la propiedad de factibilidad. Fijamos $1 > \epsilon, \delta > 0$.

Tomamos $R_1 = [a, b^*] \times [c, d]$ un rectángulo verificando $L_{\mathcal{D},f}(\mathbb{1}_{R_1}) \leq \epsilon/4$, con $a \leq b^* \leq b$.

$R_2 = [a^*, b] \times [c, d], R_3 = [a, b] \times [c, d^*], R_4 = [a, b] \times [c^*, d]$ se definen de forma análoga.

Llamando $h_R = A(S)$, $R(S) = R$ el rectángulo obtenido como resultado de aplicar el algoritmo del ejercicio para cada conjunto de entrenamiento, es claro que $P_{S \sim \mathcal{D}^m}[R \subset R^*] = 1$.

Supongamos $\forall i : R \cap R_i \neq \emptyset$. Entonces:

$$L_{\mathcal{D},f}(h_R) = P_{x \sim \mathcal{D}}[h_R \neq f] \leq P\left(\cup_i [h_R \neq f] \cap R_i\right) \leq P\left(\cup_i R_i\right) \leq 4\frac{\epsilon}{4} = \epsilon$$

La demostración acaba probando que:

$$P_{S \sim \mathcal{D}^m}[\exists i : R(S) \cap R_i = \emptyset] \leq \sum_{i=1}^{4} P[R(S) \cap R_i = \emptyset] = 4(1 - \frac{\epsilon}{4})^m \leq 4e^{-\epsilon m/4}$$

y tomando $m > \frac{4}{\epsilon} log\left(\frac{4}{\delta}\right)$.

3. Generalización aprendizaje PAC: PAC agnóstico

Hasta ahora tenemos dos problemas en la definición de PAC. Intentamos buscar una hipótesis sobre una función de verdadero etiquetado, f determinista, que por tanto no podrá asignar dos imágenes distintas al mismo punto, y además, estamos suponiendo que se cumple la propiedad de factibilidad.

Para paliar esto, podríamos considerar $\mathcal{D}$ como la distribución conjunta sobre $\mathcal{X} \times \mathcal{Y}$, y la noción de error para $h : \mathcal{X} \to \mathcal{Y}$ quedaría:

$$L_{\mathcal{D}}(h) := P_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$$

Con estos conceptos revisitados, podríamos asegurar que la hipótesis que menor error comete para $\mathcal{Y} = \{0, 1\}$ es el llamado **clasificador de Bayes**:

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & P[y = 1|x] >= 0{,}5 \\ 0 & si \quad no \end{cases}$$

Pero deseamos ir aún más allá, y generalizar la definición para una función de pérdida arbitraria.

Definición 5. ***Función de pérdida***

Dados un conjunto $\mathcal{H}$, Z y una σ álgebra de conjuntos sobre Z, se denomina función de pérdida de $\mathcal{H}$ sobre Z a cualquier función de la forma:

$$l : \mathcal{H} \times Z \to \mathbb{R}^+$$

que verifique que fijada $h \in \mathcal{H}$ arbitrario la función $l(h, \cdot)$ sea medible.

Aumiendo ya como $\mathcal{D}$ la distribución conjunta, con funciones de pérdida arbitrarias, redefiniríamos los conceptos de *error* y *error empírico* de la forma:

$$L_{\mathcal{D}}(h) := \mathbb{E}_{z\sim\mathcal{D}}[l(h,z)]$$
$$L_{S_z}(h) := \frac{1}{m}\sum_{i=1}^{m} l(h,z_i)$$

Donde los conjuntos de entrenamiento se generan a partir de una muestra aleatoria simple $S = (Z_1 \times \ldots \times Z_m)$ con $Z_i = (\mathcal{X} \times \mathcal{Y})_i \sim \mathcal{D}$

Definición 6. ***Aprendizaje PAC agnóstico***

Una clase de funciones $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ es agnósticamente PAC cognoscible respecto a $Z = \mathcal{X} \times \mathcal{Y}$ (sobre el que tenemos definida una σ álgebra de conjuntos) y a una función de pérdida $l : \mathcal{H} \times Z \to \mathbb{R}^+$ si existe una función $m_{\mathcal{H}} :]0,1[^2 \to \mathbb{N}$ y un algoritmo A verificando que si $0 < \epsilon, \delta < 1$, entonces para toda distribución $\mathcal{D}$ sobre Z ejecutando el algoritmo para un conjunto de entrenamiento $S \sim \mathcal{D}^m$, con $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ se tiene:

$$P_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \leq \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon] \geq 1 - \delta$$

El algoritmo A devuelve un elemento de $\mathcal{H}$.

Notamos desde esta definición tomando una **función de pérdida 0-1**:

$$l_{0-1}(h,(x,y)) := \begin{cases} 0 & h(x) = y \\ 1 & si \quad no \end{cases}$$

equivale a la primera definición que dimos de aprendizaje PAC si asumimos propiedad de factibilidad. Por ello no distinguiremos en el uso de uno u otro concepto, sino que se deducirá de si estamos asumiendo propiedad de factibilidad o no.

Cuando permitimos que el algoritmo A devuelva una función $h \notin \mathcal{H}$, de manera que $h \in \mathcal{H}'$ y $\mathcal{H} \subset \mathcal{H}'$ una clase de funciones donde la función de pérdida es extensible de manera natural, el aprendizaje recibe el nombre de **aprendizaje impropio**. La definición aquí dada se ha hecho para **aprendizaje propio**.

4. Condiciones suficientes para ser PAC cognoscible

Definición 7. ***Convergencia uniforme / clase de Glivenko-Cantelli***

Decimos que una clase de hipótesis $\mathcal{H}$ tiene la propiedad de ***convergencia uniforme o es de Glivenko-Cantelli****, respecto a un dominio Z, y a una función de pérdida l si existe una función $m_{\mathcal{H}}^{CU} :]0,1[^2 \to \mathbb{N}$ verificando que para todo $0 < \delta, \epsilon < 1$ y para toda distribución $\mathcal{D}$ sobre Z, si S es un conjunto de entrenamiento de tamaño mayor o igual a $m \geq m_{\mathcal{H}}^{CU}(\epsilon, \delta)$, entonces:*

$$P_{S \sim \mathcal{D}^m}[\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon] \geq 1 - \delta$$

Teorema 2. ***La convergencia uniforme es condición suficiente para ser PAC cognoscible***

Sea $\mathcal{H}$ una clase de hipótesis con la propiedad de convergencia uniforme. Entonces es agnósticamente PAC cognoscible con cualquier algoritmo ERM y complejidad muestral menor o igual al $m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right)$ dado en la definición anterior.

Demostración. Fijamos $m = m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right)$.

Fijado un conjunto de entrenamiento S_z extraído de la variable aleatoria $S = (Z_1, \dots Z_m) \sim \mathcal{D}^m$ verificando que:

$$\forall h \in \mathcal{H}, |L_{S_z}(h) - L_{\mathcal{D}}(h)| \leq \frac{\epsilon}{2}$$

Entonces, notando $\bar{h} = ERM_{\mathcal{H}}(S_z)$, para $h \in \mathcal{H}$ arbitrario:

$$L_{\mathcal{D}}(\bar{h}) \leq L_{S_z}(\bar{h}) + \frac{\epsilon}{2} \leq L_{S_z}(h) + \frac{\epsilon}{2} \leq L_{\mathcal{D}}(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} = L_{\mathcal{D}}(h) + \epsilon$$

Donde la segunda desigualdad viene desde la definición de ERM. □

Proposición 3. ***Las clases finitas tienen la propiedad de convergencia uniforme***

Sea $\mathcal{H}$ una clase de hipótesis finita, Z un dominio y sea $l : \mathcal{H} \times Z \to [a, b]$ una función de pérdida. Entonces $\mathcal{H}$ verifica la propiedad de convegencia uniforme con:

$$m_{\mathcal{H}}^{CU}(\epsilon, \delta) \leq \left\lfloor \frac{log(2|\mathcal{H}|/\delta)(b-a)^2}{2\epsilon^2} \right\rfloor + 1$$

Lema 1. ***Desigualdad de Hoeffding***

Sean $(X_1, \dots X_m)$ una muestra aleatoria simple de una variable X, $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ con $E[\bar{X}] = \mu$ y $P[a \leq X_i \leq b] = 1, i = 1, \dots m$. Entonces para todo $\epsilon > 0$

$$P\left[\left|\bar{X} - \mu\right| > \epsilon\right] \leq 2e^{-2m\left(\frac{\epsilon}{b-a}\right)^2}$$

Demostración. Sea $\mathcal{H}$ una clase de hipótesis finita.

Fijamos $0 < \delta, \epsilon < 1$. Necesitamos encontrar $m \in \mathbb{N}$ verificando:

$$P_{S \sim \mathcal{D}^m}[\exists h \in \mathcal{H} | L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] < \delta$$

Partimos de la siguiente desigualdad, que usaremos más adelante, obtenida por subaditividad:

$$P[\exists h \in \mathcal{H} | L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] \leq \sum_{h \in \mathcal{H}} P[|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon]$$

Fijamos $h \in \mathcal{H}$.

Dado un conjunto de entrenamiento $S_z = (z_1, \dots z_m)$, recordamos que $L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}}[l(h, z)]$ y que $L_{S_z}(h) = \frac{1}{m}\sum_{i=1}^{m} l(h, z_i)$

Donde $z_i \sim \mathcal{D}$ y por tanto $\mathbb{E}_{S \sim \mathcal{D}^m}[L_S(h)] = \mathbb{E}_{z \sim \mathcal{D}}[l(h, z)] = L_{\mathcal{D}}(h)$. Además, llamando $X_i = l(h, Z_i)$, por ser $S = (Z_1, \dots Z_m)$ m.a.s que genera los conjuntos de entrenamiento, se tiene que las X_i son independientes e idénticamente

distribuidas, con $P[a \leq X_i \leq b] = 1$. Estamos en condiciones de aplicar la desigualdad de Hoeffding.

Por tanto:

$$P\left[\left|\frac{1}{m}\sum_{i=1}^{m} X_i - L_{\mathcal{D}}(h)\right| > \epsilon\right] = P[|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] \leq 2e^{-2m\left(\frac{\epsilon}{b-a}\right)^2}$$

Y por tanto:

$$P[\exists h \in \mathcal{H} | L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] \leq |\mathcal{H}| 2e^{-2m\left(\frac{\epsilon}{b-a}\right)^2}$$

Despejando m para que $|\mathcal{H}| 2e^{-2m\left(\frac{\epsilon}{b-a}\right)^2} < \delta$ llegamos al resultado buscado. □

Recordemos hasta ahora el resultado que habíamos obtenido era su carácter PAC cognoscible, donde agnósticamente PAC cognoscible y cognoscible con funciones de pérdida 0-1 era un término equivalente. El teorema que enunciamos a continuación, deducible a partir del teorema sobre el caracter agnóstico - PAC cognoscible de clases de funciones con propiedad de convergencia uniforme, en particular las finitas, generaliza el resultado para cualquier funciones de pérdida acotada.

Teorema 3. ***Las clases finitas son agnósticamente PAC cognoscible***

Sea $\mathcal{H}$ una clase de hipótesis finita, Z un dominio y sea $l : \mathcal{H} \times Z \to [a, b]$ una función de pérdida. Entonces $\mathcal{H}$ es PAC cognoscible con complejidad muestral:

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2log(2|\mathcal{H}|/\delta)(b-a)^2}{\epsilon^2} \right\rceil$$

Demostración. Es trivial desde el anterior teorema y el hecho de que convergencia uniforme implica ser agnósticamente PAC cognoscible □

5. Equilibrio error-varianza

Veamos que dado un algoritmo de aprendizaje no puede ser el óptimo para aprender todas las distribuciones.

Damos un lema previo, la desigualdad de Markov:

Lema 2. ***Desigualdad de Markov***

Dada una variable aleatoria Z no negativa. Entonces para todo $a \geq 0$

$$P[Z \geq a] \leq \frac{\mathbb{E}[Z]}{a}$$

Demostración. Sea f la función de densidad de Z.

$$aP[Z \geq a] = \int_a^{+\infty} af(z)dz \leq \int_a^{+\infty} zf(z)dz \leq \int_0^{+\infty} zf(z)dz = \mathbb{E}[Z]$$

□

Teorema 4. ***Teorema de No Free Lunch***

Sea A cualquier algoritmo de aprendizaje para clasificación binaria con respecto a la función de pérdida 0-1 sobre el dominio $\mathcal{X}$. Sea un conjunto de entrenamiento de tamaño $m < |\mathcal{X}|/2$. Entonces existe una distribución $\mathcal{D}$ sobre $\mathcal{X} \times \{0, 1\}$ verificando:

1. *Existe una función $f : \mathcal{X} \to \{0, 1\}$ con $L_{\mathcal{D}}(f) = 0$*
2. *$P_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(A(S)) \geq 1/8] \geq 1/7$*

Demostración. Sea un conjunto de entrenamiento (consideramos un conjunto y no una secuencia) de tamaño $2m$, C. Hay $T = 2^{2m}$ posibilidades de etiquetado del conjunto, esto es, T posibles hipótesis, $f_i : C \to \{0, 1\}$, que vamos a extender a $\mathcal{X}$ llamándolas $\bar{f}_i$ de forma que $\bar{f}_{i|C} = f_i$ y $\bar{f}_i(x) = 0 \quad \forall x \in \mathcal{X} \setminus C$. Vamos a tomar para cada una de ellas una distribución $\mathcal{D}_i$ definida sobre $\mathcal{X} \times \{0, 1\}$ definida por:

$$\forall (x,y) \in \mathcal{X} \times \{0,1\} \qquad P_{Z\sim\mathcal{D}_i}[Z=(x,y)] = \begin{cases} 1/|C| & x \in C, y = f_i(x) \\ 0 & si \quad no \end{cases}$$

Claramente $L_{\mathcal{D}_i}(f_i) = 0$. Tenemos distribuciones de probabilidad que sólo asignan toda la masa de probabilidad a la marginal en $\mathcal{X}$ al conjunto C.

Vamos a probar que:

$$\exists i \in \{1, \dots T\} : \mathbb{E}_{S\sim\mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] \geq \frac{1}{4}$$

Fijamos $i \in \{1, \dots T\}$. Hay $k = (2m)^m$ posibles tuplas de tamaño m, $S_j, j = 1, \dots k$ tomadas desde C. Siendo $S_j = (x_1, \dots x_m)$ notamos $S_j^i = ((x_1, f_i(x_1)), \dots, (x_m, f_i(x_m)))$. Cada S_j tiene la misma probabilidad de ser nuestro conjunto de entrenamiento (extracción de m valores con reemplazamiento desde el conjunto C), verificándose:

$$\mathbb{E}_{S\sim\mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] = \frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i))$$

Recordando que hemos llamado $k = (2m)^m$, $T = 2^{2m}$, se tiene:

$$\begin{aligned} max_{i\in\{1,\dots T\}} \frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) = \\ &= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \geq \\ &\geq min_{j\in\{1,\dots k\}} \frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \end{aligned}$$

Además fijado $j \in \{1, \dots k\}$:

Sean $v_{r_{i=r}}^{p}$ los elementos de C no presentes en el conjunto de entrenamiento S_j. Claramente, como $|C| = 2m$ y $|S_j| = m$ y puede tener elementos repetidos, $p \geq m$

$$L_{\mathcal{D}_i}(A(S_j^i)) = \frac{1}{|C|} \sum_{x \in C} \mathbb{1}_{[A(S_j^i)(x) \neq f_i(x)]} = \frac{1}{2m} \sum_{x \in C} \mathbb{1}_{[A(S_j^i)(x) \neq f_i(x)]}$$

Por tanto:

$$\begin{aligned} \frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) &\geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2m} \sum_{x \in C} \mathbb{1}_{[A(S_j^i)(x) \neq f_i(x)]} \geq \\ &\geq \frac{1}{2p} \sum_{r=1}^{p} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} \geq \\ &\geq \frac{1}{2} min_r \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} \end{aligned}$$

Como dado un v_r cualquiera, $v_r \notin S_j$, y existen $f_i, f_{i'}$ que se diferencian justo en el elemento v_r, uno coincidirá con el valor en v_r de $A(S_j^i) = A(S_j^{i'}$ y otro no:

$$\frac{1}{2} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} = \frac{1}{2} \frac{1}{T} \frac{T}{2} = \frac{1}{4}$$

Y uniendo toda esta información:

$$max_{i \in \{1, \ldots T\}} \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) \geq \frac{1}{4}$$

Sea $k = argmax_{i \in \{1, \ldots T\}} \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i))$

Si $\mathcal{D} = \mathcal{D}_k$ cumple la parte 2 del enunciado del teorema, es nuestra distribución buscada, y como función buscada en el apartado 1. podemos tomar $f = f_k$

Como $L_{\mathcal{D}}(A(\cdot))$ puede ser vista como una variable aleatoria donde $S \sim \mathcal{D}^m$ y que toma valores en $[0, 1]$, tenemos que tomando $Z = 1 - L_{\mathcal{D}}(A(\cdot))$, $a = \frac{7}{8}$ en el lema previo llegamos a:

$$P_{S\sim\mathcal{D}^m}\left(\frac{1}{8} \geq L_{\mathcal{D}}(A(S))\right) \leq \frac{3}{4} \cdot \frac{8}{7} = 24/28$$

donde $\mathbb{E}(Z) = \mathbb{E}(1 - L_{\mathcal{D}}(A(\cdot))) = 1 - \mathbb{E}(L_{\mathcal{D}}(A(\cdot))) \leq \frac{3}{4}$

Es decir:

$$P_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}\right) \geq \frac{4}{28} = \frac{1}{7}$$

□

Como consecuencia del teorema, podemos decir que no hay un algoritmo de aprendizaje óptimo para todas las distribuciones, puesto que para una dada por el resultado del teorema, el algoritmo ERM con $\mathcal{H} = \{f\}$ aprendería mejor.